



INSTITUTE FOR DEFENSE ANALYSES

**Designing a Qualitative Data Collection Strategy
(QDCS) for Africa –
Phase I: A Gap Analysis of Existing Models,
Simulations, and Tools Relating to Africa**

Ashley N. Bybee, Project Leader
Dominick E. Wright

June 2012

Approved for public release;
distribution is unlimited.

IDA Document D-4629

H12-000748



The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract DASW01-04-C-0003, Task AI-55-3061.0.0, "Designing a Qualitative Data Collection Strategy for Africa," for the Assistant Secretary of Defense for Research and Engineering. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Approved for public release; distribution is unlimited.

Copyright Notice

© 2012 Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper D-4629

**Designing a Qualitative Data Collection Strategy
(QDCS) for Africa –
Phase I: A Gap Analysis of Existing Models,
Simulations, and Tools Relating to Africa**

Ashley N. Bybee, Project Leader
Dominick E. Wright

Executive Summary

This study is sponsored by the Rapid Reaction Technology Office (RRTTO) in the Office of the Deputy Assistant Secretary of Defense for Rapid Fielding (DASD/RF), for the Assistant Secretary of Defense for Research and Engineering (ASD/R&E). OSD recognizes the value models and simulations (M&S) and other computational tools have to support decision-making at the strategic, operational, and tactical levels, while also aiding in intelligence analysis and information and contingency operations. The value of M&S, however, is contingent on the availability of high-quality data input to the system. Their performance is greatly improved with the use of high-quality, validated data that provide a more accurate picture of the social, cultural, political, and even economic landscapes in various African regions.

This study will ultimately present a Qualitative Data Collection Strategy (QDCS) that addresses “gaps” where insufficient qualitative data exist for the African continent. This will allow the U.S. Government (USG) to improve its qualitative data collection efforts and ensure that the most accurate and valid data are captured and used in its M&S. Because this study focuses exclusively on the African continent, IDA sees the Africa Command (AFRICOM) as a primary beneficiary of this

research as well as its components and other organizations within the USG with activities in Africa.

The long-term goal of this line of inquiry is to enable more accurate social science modeling, recognizing that policy-makers should not expect M&S performance to be so accurate as to serve as a predictor of probable eventualities. On the contrary, social science M&S are intended to characterize complex socio-cultural-economic-political dynamics rather than to identify the result of a given scenario. In making such characterizations, they reveal possible interactions, inform the decision-making process, and provide a point of departure for further discussion, research, and analysis. Moreover, a comprehensive survey of data requirements and accompanying qualitative data collection strategy will ensure the most efficient allocation of resources and avoid duplication of data collection efforts.

This document summarizes the findings of the IDA team’s first phase of research, which was a survey of those models, simulations, and relevant tools (MS&T) currently being used to analyze the African continent. It identifies the most pressing gaps in data either expressed by those MS&T designers/users or identified by IDA. It also captures the most significant challenges or obstacles to effective data collection, which are relevant as IDA develops its QDCS.

Contents

The Briefing	1
Premises.....	3
IDA Approach and Objective	5
Scope and Methodology	7
Incorporating Social Science Data into MS&T	9
MS&T Survey	11
Findings: Data Gaps	13
Accurate Representation of Relative Effects	15
Data Characterizing the Analytical Framework	19
Additional Gaps.....	21
Challenges of Collecting/Analyzing Qualitative Data from Africa	25
Next Steps.....	29
Appendix: MS&T Survey	31

Designing a Qualitative Data Collection Strategy (QDCS) for Africa

***Phase I: A Gap Analysis of Existing Models, Simulations, and
Tools Relating to Africa***

June 2012

Authors:

Dr. Ashley Bybee, Project Lead

Dr. Dominick Wright

The dearth of both qualitative and quantitative data on Africa is well documented and a result of limited resources on such a massive, complex continent. Due to the limited data collection by indigenous research organizations, most academic and policy-related research relies heavily on qualitative data produced by Western-educated social scientists, anthropologists, and political scientists. Since they often lack an African perspective, which can militate against cultural bias inherent in a Western approach, it is difficult to gauge the pertinence, accuracy, and explanatory power of much qualitative data in the African context.

Moreover, the sensitive nature of many “taboo” research areas – including sexual behavior, religious practices, lifestyle habits, drug consumption, as well as important security issues such as illicit trafficking (e.g., Small Arms/Light Weapons, Weapons of Mass Destruction, drugs, humans), and the nexus of each with terrorism – has inhibited meaningful discussion in these areas and the elicitation of native perceptions of these issues. As a result, these qualitative data points are currently, and understandably, absent for Africa. Given that the performance of modeling and simulation (M&S) and their ability to characterize the environment are based on these and related qualitative data inputs, the concern is that such performance is impaired, or at the very least could benefit greatly from improved data collection.¹

¹ Ongoing creation of the J-50 Cell at AFRICOM, whose primary mission is the collection, storage, analysis, and dissemination of civil information, attests to the importance of the issue addressed in this report from the perspective of the Command.

From a data user’s perspective, there are additional concerns relating to the supply, format, and quality of qualitative data that affect their ability to be used in MS&T. A common concern voiced by researchers is that it has been difficult to identify subject matter experts (SMEs) who can provide reliable, high quality data. Another issue is consistent access, which many data users lack. For example, they may receive one or two qualitative data sources that represent valuable “snapshots” at a given point in time, but they do not have access to those same data over time. This absence of sufficient “time series” data makes it difficult or even impossible to understand overarching trends that are so critical for researchers to determine causalities and the relationships with other variables examined. Moreover, without some indigenous insight into local geo-political and environmental conditions, it is inordinately more difficult for outsiders to know how relevant and valid data are over time, or how applicable the data may be across a span of operational contexts. Finally, although there is a DoD data standard within the context of M&S (DoD Directive 5000.59, 2007), adherence to its prescriptions is variable. Combine this with the fact that data used for alternate purposes have alternate standards as well,² and the problem of regularity in critical data features, such as format, becomes clear. As a result, qualitative (and quantitative) data lack consistency in critical features, such as format and unit of measurement, which makes it difficult to achieve seamless inputs into MS&T.

² For example, geo-spatial and motion imagery data have DoD standards.

**Dearth of Qualitative Data (QD)
from the African AOR**

- No consistent access to high quality data
- No reliable way to identify, vet, and use SMEs who can supply data

**No sense for the quality of existing
QD or if it has sufficient explanatory
power in the African context**

- No sense for how conditions (e.g., environmental, geo-political) determine data relevancy over time
- No systematic process for identifying data suitable across operational contexts and data limited to specific contexts
- Current supply of data is not consistently configured according to a single standard and existing DoD requirements

***Insufficient data produce “gaps” impairing the
performance and utility of M&S analyzing Africa***

IDA is approaching this project in three phases. The first phase is a survey of the existing MS&T currently being used to analyze the African continent. IDA did not have any externally applied constraints or limitations on the projects to be surveyed as part of this phase, and sought rather to develop a keen sense of what capabilities are most used and most desired by the community of Africa analysts. Toward that end, IDA surveyed MS&T currently used by USG organizations and some in development in academia. IDA found that the number of modeling and simulation projects was far smaller than anticipated, perhaps a testament to the lack of data with which to model and/or simulate African phenomena. For each of the projects, IDA interviewed the designers/users of the MS&T and sought the following information: types/sources of qualitative data used, data collection/validation methodologies, assessment of data quality, format of data, challenges to collection/analysis, and gaps in qualitative data.

The second phase of the task involves engagement with African scholars and Africa-based researchers. IDA views this step as a critical feature of a data collection strategy, because it takes into account indigenous insights into African issues. While

U.S.-based researchers and designers of MS&T have clear data requirements for their systems, such data may not have the most explanatory power in the African context. As a result, they may be analyzing data that do not reveal new insights or shed new light on emerging trends. Soliciting input from Africans on capturing what they perceive to be the most salient information to explain certain phenomena, while identifying emerging trends that might not be on Americans' radar, represents a strategic investment with immediate and long-term returns. Moreover, including Africans as active participants in this phase of strategizing will better position the USG and research institutions to engage Africans on issues of mutual concern in the future and cultivate long-term partnerships that yield new data for both the U.S. and its African partners.

The third phase of this task will entail the drafting of the final report and the development of the final strategy. The resultant Qualitative Data Collection Strategy (QDCS) should serve the needs of the MS&T communities relating to Africa, while ensuring that data requirements are attuned to the interests of African partners.

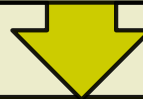
This report documents IDA's findings from Phase I.

Phase I: “Gap” analysis of MS&T relating to Africa

Reliability: Is data available? On a consistent basis? Only some of the time? Does it consistently capture the same information?

Validity: Is data accurate? Does it reflect the concept it purports to?

Structure: Is the data structured in a format conducive for use in M&S?

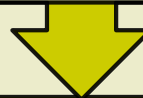


Phase II: Solicit input from African partners

Are gaps identified by Americans the only ones? Are they relevant?

Identify additional data points from African perspective

Including Africans as active participants in data collection efforts ensures that data is as accurate and as valid as possible



Phase III: Strategy Development

Design a Qualitative Data Collection Strategy (QDCS) for acquiring missing data relevant to the study and analysis of Africa

Objective: Design a QDCS that will service the needs of those who use models, simulations and tools relating to Africa

A model is a simplified representation of a system at some particular point in time or space intended to promote understanding of the real system. A simulation is the manipulation of a model in such a way that it operates on a time or space continuum, thus enabling one to observe possible interactions that would not otherwise be apparent because of their separation in time or space. Other computational tools, such as those that assist in the collection, aggregation, organization, and illustration of data, were not initially included in the scope of this study, but were added later to accommodate an analysis of Serengeti, which is currently in use by AFRICOM.³ The main MS&T targeted by the IDA team were those currently being used by the USG for the analysis of Africa. Because the actual number of MS&T used by the USG was not as high as expected, however, IDA broadened the scope of the study to include MS&T used in academia, since their qualitative data gaps are also helpful data points. Even with the expanded scope, there still remain relatively few Africa-focused M&ST.

DoD Memorandum 5000.59-M defines qualitative data as, “a data value that is a non-numeric description of a person, place, thing, event, activity, or concept.” Extending this definition, the Oxford Dictionary of Statistical Terms describes it as an attribute and states:

“A qualitative characteristic of an individual, usually employed in distinction to a variable or quantitative characteristic. Thus, for human beings sex is an attribute but age is a [quantitative] variable. Often attributes are dichotomous, each member of a population being allotted to one of two groups according to whether he or she does or does not possess some specified attribute; but manifold classification [i.e., organization across multiple categories] can also be carried out on the basis of attributes...” (p. 18)

³ Due to scheduling conflicts, meeting with AFRICOM was not possible before completion of this deliverable. IDA still seeks a meeting with the SERENGETI team as well as other consumers of qualitative data in the Command. Findings from a future meeting will be provided to the sponsor in a separate memo.

“[Quantitative data], in contrast to qualitative data, should relate to data in the form of numerical quantities such as measurements or counts.” (p. 328)⁴

For the purpose of this study, *qualitative data* refers to data that, regardless of format (numeric or text), are inherently subjective in nature and therefore require some interpretation when coding. Polling data measuring public opinion, unstructured data gathered from focus groups, or data gathered through various anthropological approaches such as participation observation (the study of a society through participation in its environment) are all examples of qualitative data. While there is also a well-documented lack of quantitative data from Africa, this study focuses on the more complex data that can contribute to an in-depth understanding of human behavior and the reasons that govern such behavior. Qualitative data can help to answer questions about the “why” and “how” of decision making and certain phenomena, not just “what,” “where,” or “when.”

The methodology employed for this first phase of research included an extensive literature review of known reports and articles on the subject of M&S in an effort to identify those applicable to Africa and suitable for further study. IDA reviewed DoD’s M&S Catalog⁵ and, with guidance from OSD, contacted several individuals leading projects relevant to the study. Additional recommendations and points of contact for applicable MS&T were provided, which enlarged IDA’s pool of interviewees to a sufficient sample size. IDA continues to engage with MS&T owners and data providers, as identified, for inclusion in the study.

⁴ *The Oxford Journal of Statistical Terms*. Ed. Yadolah Dodge. New York: Oxford University Press, 2003.

⁵ Available at: <https://mscatalog.osd.mil/intro/index.aspx>

- **Scope**

- Models, simulations, and relevant tools currently in use by the USG, or with the potential to be used by the USG, for analysis of the African AOR.
- Qualitative Data: Regardless of format (numeric or text), data that are inherently subjective in nature, based upon individual perspectives and requiring interpretation when coding.

- **Methodology**

- Review of literature pertaining to M&S
- Interviews with MS&T owners, users, and data providers

All data – quantitative and qualitative – must be reviewed and adequately standardized before using them to conduct rigorous analysis.⁶ Qualitative data present a unique set of challenges for social scientists, particularly when the data must be coded for incorporation into MS&T. The diagram here represents the stages that qualitative data must pass through before they can be used in meaningful analysis. All the MS&T IDA surveyed were processed in this way.

From the outset, collection of qualitative data can be challenging, particularly where they are sought in austere and/or denied environments, such as insecure regions of Somalia or where corruption and government bureaucracy impedes effective collection. Because qualitative data collection can be very costly and time-consuming, researchers must know the exact data points they wish to collect, which entails a deep understanding of the most salient indicators to track in different cultures. Collection must also account for conditions that may skew samples though response biases.

Collected data must then be translated from unstructured information into data that are formatted and adequately “tagged” with sufficient metadata so that they are useful for operators and analysts.

Translated data must then be organized into a taxonomy that is logical and facilitates use across a range of domains,

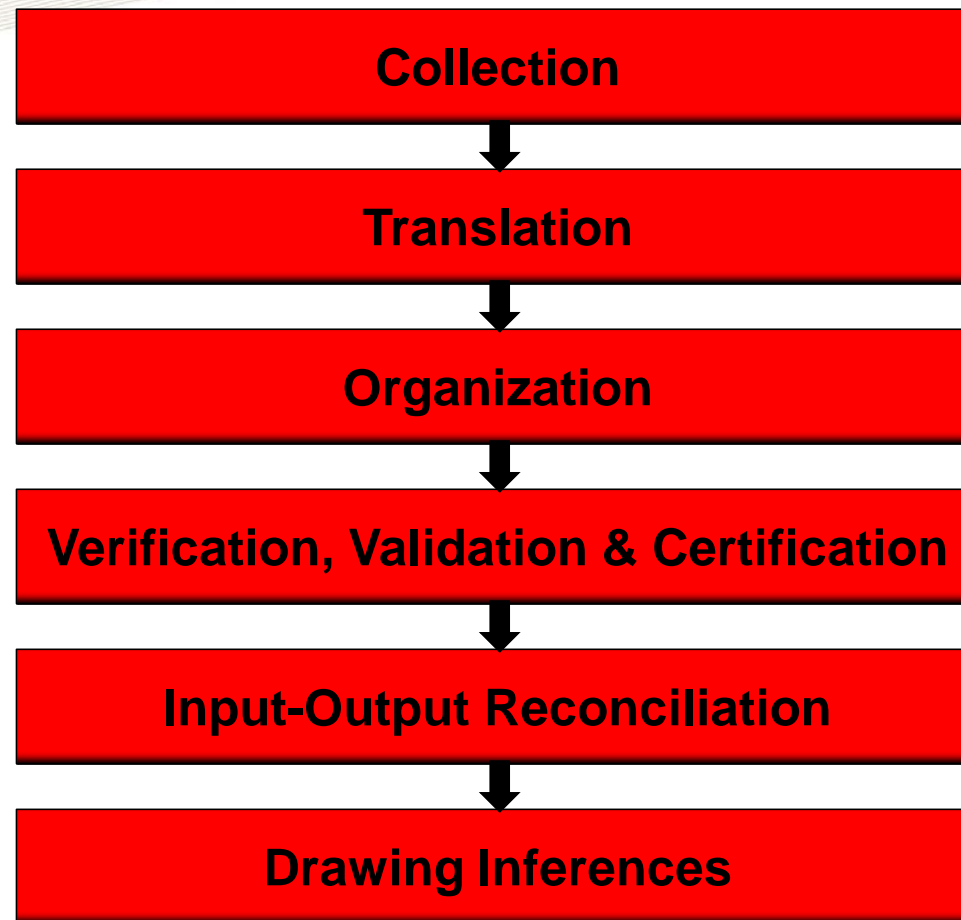
including academia, scientific, military, and other end user communities.

Perhaps the greatest challenge of all when using qualitative data is verification, validation, and accreditation. As published in the DoD Directive (DODD) 5000.59, data verification ensures that data accurately represent the developer’s conceptual description and are transformed and formatted properly. Validation is the assessment of data by SMEs or comparison of data to known or best-estimate values. This ensures the data accurately represent real life. Certification is the recognition that the data are acceptable for use for a specific purpose (such as for use in a model or simulation).

Then, qualitative data (the input) might need to be modified in order to “fit” into the model or simulation (the output). This often occurs when the sampling unit does not correspond with that of the receiving model or simulation, such as country-level nutrition rates versus individual-level observations of health.

Finally, correctly interpreting qualitative data requires a certain level of familiarity with the subject matter. M&S users must recognize that the range of results might be influenced by collection conditions or other externalities. While the qualitative data are still highly valuable, drawing inferences requires careful attention to all aspects of the data.

⁶ While it is possible to apply rigorous methods to inaccurate data, it goes without saying that the potential for analytic gains looms large over the prospect of substituting accurate for fallacious data.



The slide on the opposite page lists the models, simulations, and tools that IDA surveyed for the purpose of this study. Their names and the organization responsible for their development are

also included. A more comprehensive description of each, including name, producer, type, purpose, and inputs are described in detail in Appendix A.

Name or Description	Type	Organization
Competitive Influence Game (CIG)	Simulation (Independent & Federated)	John Hopkins University Applied Physics Laboratory (APL)
Composite Vulnerability Map	Tool (Web-Based)	Climate Change and African Political Stability Program (CCAPS), University of Texas
Cultural Geography (CG)	Model (ABM)	TRADOC Analysis Center (TRAC)-Monterey
Geospatial Information Awareness / Infection Disease (GIA/ID)	Model (Analytic)	Naval Research Lab (NRL)
HOA-Viewer	Tool (Web-Based)	Department of State (DoS), Humanitarian Information Unit (HIU)
Information Velocity 2.0 (IV2)	Tool (Web-Based)	Office of the Secretary of Defense, Science and Technology (OSD-S&T)
RiftLand	Model (ABM)	Center for Social Complexity, George Mason University
N/A	Model (Statistical) and Tool (Web)	Naval Postgraduate School (NPS), Operations Research Department
*Serengeti	Tool	AFRICOM/Undersecretary of Defense (Intelligence) (USD(I))

***Meeting requested**

Findings: Data Gaps

The ability to accurately represent the relative effect of a given action remains a challenge for all M&S where qualitative methodologies are employed. Unlike quantitative data, which are useful for statistical analysis, qualitative data may need to be quantified – a process inherently requiring certain assumptions to be made. Due to the blurry lines dividing qualitative (ordinal or categorical) variables, decisions regarding thresholds or definitions are often arbitrary and subject to individual interpretation. Whereas statistical analysis of quantitative data produces comparatively robust coefficients describing levels of change in an outcome variable from commensurate changes in a given input, qualitative data cannot calculate this coefficient so easily.⁷

Graphs in the diagram use a notional, linear regression formulation to capture relative relationships between variables. Theoretical perspectives in the literature of the relationship between communal hardship and popular support for violent extremist action (e.g., militancy) commonly assert that as popular wellbeing decreases, support for extremism increases. Although clear in its broad strokes, the perspectives do not reveal the exact nature of the relationship, especially on a locale-by-locale basis (e.g., according to country, districts within countries, and communities within districts). More importantly, analysts, planners, and others interested in such relationships lack information on the influence that developmental assistance has on wellbeing. The diagram illustrates the quantities summarizing these relationships as β_1 , the effect that developmental projects (e.g., drilling a borehole to improve public access to water), and β_2 , changes in levels of popular support for extremism conditioned on changes in popular wellbeing. Represented in simple,

bivariate (or two-variable) terms here, factors governing each relationship are in truth potentially complex. For instance, understanding the number of individuals likely to derive net benefits from the drilling of a borehole in a community requires knowing more than the level of water scarcity (something perhaps gleaned from imagery data); it also requires knowing other community features, such as:

- The existence of conflict in the community
- Who controls access to the borehole
- What effect an additional borehole will have on communal and inter-communal relations.

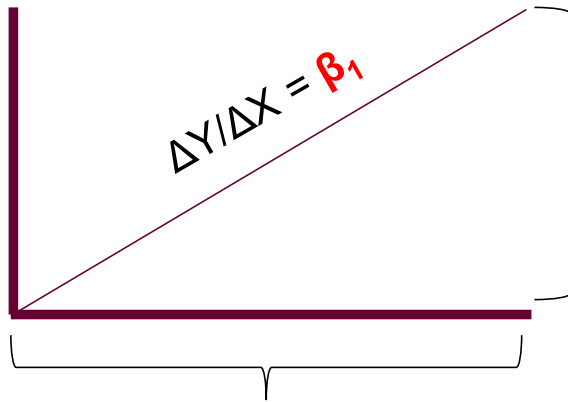
Because each region, community, or other unit of analysis is unique, making a one-size-fits-all approach inappropriate, the collection of relevant qualitative data is a resource-intensive process. In terms of modeling the effects of such an action, designers must know how the action will affect the distribution of results, i.e., how an action will affect the mean and variance in a normal distribution or, more generically speaking, how the action will affect the parameters of the distribution describing the range of possible outcomes.

In lieu of adequate qualitative data to inform this calculus, the M&S that IDA surveyed generally defined these rates based on the designers' own judgment and reasoning. For example, in the GMU-developed model RebeLand with a variant called RiftLand that will model humanitarian crises in East Africa, issues enter the environment with a user-defined onset rates, a log-normal decay rate,

⁷ Remedies to this situation include not only larger sample sizes but also improved qualitative understandings drawn from focus groups, unstructured surveys, and so forth that improve the identification, location, and contextualizing conditions of relevant thresholds.

Accurate Representation of Direct and Indirect Relative Effects

E.g. How effective will development assistance be on increasing popular wellbeing? As popular wellbeing increases, what is its effect on support for militancy?



ΔX
(Action)



X = Number of donor-built boreholes
(Development Assistance)

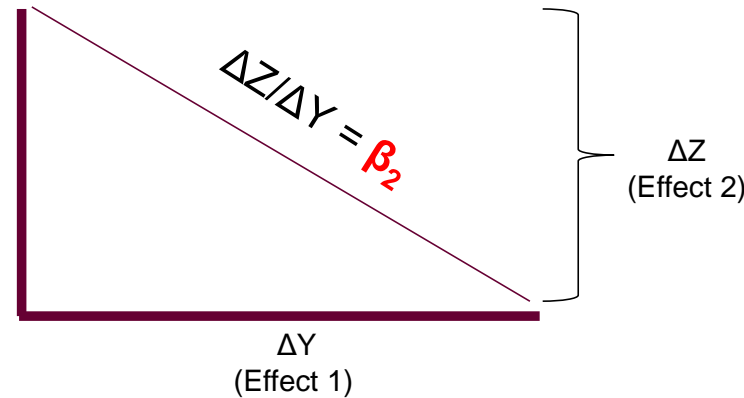


Y = Percentage of the population
with access to potable water
(Popular Wellbeing)

ΔY
(Effect 1)



Z = Popular support for extremism



and a power-law distributed magnitude.⁸ This allows users to define the level of stress a government will probably face in a given simulation run. In the army's Competitive Influence Game (CIG), causal relationships exist, but the size of the effect, or the degree to which a given action contributes to an outcome, is based on expert judgment following a detailed literature review of academic materials and discussions with SMEs. Because neither is empirically driven (i.e., derived from systematic analysis of observed values), the effects of an action are not as accurate as they could be.⁹

The inability to accurately model or simulate relative direct and indirect effects of a given action represents a shortcoming in many M&S. Nonetheless, this capability is a critical component of the calculation for modeling/simulating interactions with any level of accuracy and reality. In the African context, possible qualitative data requirements to inform this calculus might include:

- Effectiveness of insurgent recruitment
- Effectiveness of information operations
- Propensity of a given population to take risks
- The degree to which a population adheres to cultural norms.

Consider, for example, the assessment when thinking of rates of adherence to cultural norms, such as what percentage of the population are *devout* followers of Islam and can be expected to engage in *all* related activities versus more liberal or fair-weather Muslims who are not as committed to all Islamic practices.

Similarly, the strength of alliances can be measured on a spectrum that could impact whom the U.S. military decides to leverage for collaboration in order to have the greatest impact. To model these examples, one must have sufficient qualitative data to empirically derive the distribution of effects. Once identified, models must ask whether these data are scalable and be aware of what variables may cause them to change. Once resolved, M&S can be more effective in tracking and assessing MOEs.

⁸ Claudio Cioffi-Revilla and Mark Rouleau, "MASON RebeLand: An Agent-Based Model of Politics, Environment, and Insurgency" *International Studies Review* 12, 31-52, 2010.

⁹ This should not be interpreted as a critique of CIG, which due to time constraints, limited resources and greater emphasis on other aspects of the game precluded such a time-consuming process. The designers of CIG readily admit the use of qualitative data in the game is imperfect and are looking to improve it in future cycles.

Modeling African societies and other socio-cultural phenomena generally suffers from poor insight into the relationships among actors, influences, authorities, values, ideologies, and other aspects of the environment in question. An analytical framework that accurately maps connections among these components and the values of these connections (e.g., direction and level of influence) is often absent. Hence, it is a major gap. This diagram illustrates the point with examples of some components that might comprise such an analytic framework.

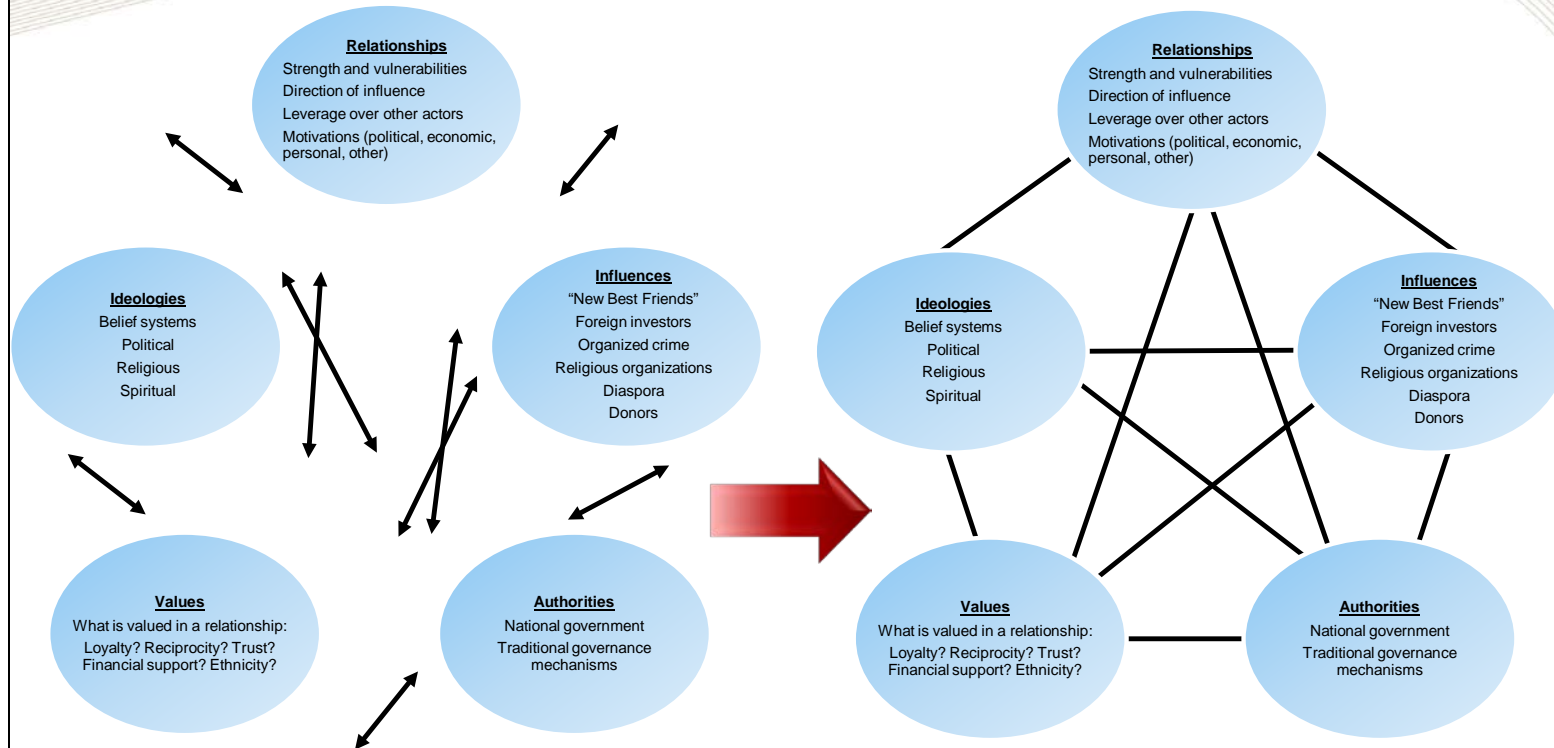
Ideally, M&S would have enough validated data to characterize each modeled relationship, accurately portraying how a given scenario might unfold. The collection of data at this level of detail, however, is difficult for multiple reasons. Not only is it costly in terms of resources and time, but the dynamic nature of the socio-cultural trends and phenomena within African communities makes it difficult to be confident that data are always accurate. The more realistic alternative described here is a validated theoretical framework that maps the most basic connections and provides analysts with some insight on relevant subject areas, such as the influence of new “best friends,”¹⁰ the role of traditional versus national governance mechanisms, and identification with as well as membership in, for instance, religious/animist/spiritual sects. The nature of these relationships is a gap IDA identified whose closure could assist in the

enhanced modeling of the complex, dynamic African environment.

In CIG, designers provided a basic description of Diplomatic, Intelligence, Military and Economic (DIME) inputs and Political, Military, Economic, Social, Infrastructure, and Information (PMESII) outputs, which relied heavily on SME input from sociologists, cultural anthropologists and demographers to characterize relationships. SMEs provided qualitative data that described ethnic groups, socio-demographics, and other attributes relevant to an initial understanding of human terrain. The resultant JIPOE¹¹ description was an outlining of “initial conditions” necessary for contextualizing the exercise, but the simulation primarily motivated “game play” and stimulated discussion, rather than helping to develop actual plans of action, because the critical, relational aspects of its mechanics had no empirical anchors. Time and resource constraints are the primary cause for this gap.

¹⁰ “New Best Friends” is a term that refers to non-African countries that have, within the last five to seven years, exhibited interest in increasing economic, diplomatic, and even military ties with African countries. Examples include Turkey, Iran, U.A.E., Saudi Arabia, and some other Arab states.

¹¹ In military terms, the practice of gathering information to build an understanding of a given environment is called a “Joint Intelligence Preparation of the Operational Environment” (JIPOE). The JIPOE process is a critical one and ideally produces a list of the most important variables. The mapping of these variables to one another to assist in analysis of relative effects, however, is often missing.



Time Series Data

A significant deficiency in much qualitative data from Africa is that a constant and sustainable flow of reliable data typically does not exist. Several “snapshots,” i.e., data points from a specific point in time, could exist and be useful, but without capturing repeated observations in the form of time series data, it is impossible to track trends. This is especially important in survey data to capture changing attitudes over time as influencers, external forces, and other components within the system change. Time series data might also serve as a valuable validation tool, where multiple variables can be compared over time, allowing analysts to better understand the entire system of phenomena being analyzed. Although long-term collection requires considerable investment of resources, time, and infrastructure, it is a vital requirement since M&S need access to a constant flow of valid, reliable, affordable, timely, and properly formatted data for use by analysts.

Sub-national Data

National-level data are often available for some African countries, such as socio-economic indicators collected from national censuses or governance scores calculated by research institutes. A frequently cited gap in data for Africa (qualitative as well as quantitative), however, is sub-national and sub-regional data, e.g., data at the provincial, district, and village levels. Because colonially drawn national borders typically don’t reflect a homogeneous group within, socio-cultural datasets at the national level often exhibit a wide range of responses that are

unhelpful and potentially misleading when analyzing the local population and environment. Data collected at the appropriate sub-national level have much greater utility in M&S in that researchers can identify the population in question and collect the pertinent data necessary to measure and understand local phenomena.

Perception Data

There are numerous survey and polling projects underway in Africa that reflect the recognition by the USG that data on perceptions, attitudes, and public opinion are highly valuable data points. The number of surveys conducted in Africa by both the USG and foreign research institutions (including many African institutions) is an encouraging trend. Nonetheless, the M&S community continues to identify new requirements for public opinion data, such as perceptions of “new best friends,” perceived strength of those relationships and depictions of micro-level, internal dynamics.

Geographic Gaps

Qualitative data exhibit many obvious geographic gaps, i.e., countries or regions for which few to no data are available. Among these geographic gaps are areas where access is limited because of difficult terrain (remote and desolate areas), countries where there are political impediments to effective data collection (authoritarian regimes that oppress free speech such as in Eritrea), or areas where insecurity poses too great a risk for data collectors (such as Somalia or Mali.)

- Time series data
- Sub-national data
- Perception data
- Geographic gaps

Baseline Data

The general dearth of data in Africa is also true of baseline data, i.e., the initial dataset that serves as a basis for comparison with subsequently acquired data. Without baseline data, however, it is impossible to develop effective metrics and evaluate progress. As a result, some researchers have shifted focus to analysis of long-term phenomena such as climate change, examining case studies over the course of the next 20 years. This long-term approach gives them better perspective and the ability to track new issues that may be more appropriate in determining new measures of effectiveness (MOEs).

Recent Data

There are many regions and subject areas in Africa where recent data are unavailable, either due to insecurity (such as in Somalia) or because there are no sustained data collection efforts in place and previous collections have been largely serendipitous. As with time series data, the lack of recent data is another major gap in many African countries and regions that makes timely analyses impossible.

Census Data

Census data are available for some African countries that have achieved a certain level of technical capacity for data collection. A common complaint among the M&S community, however, is that in many countries, censuses may not be publicly unavailable, they may be unreliable (due to poor collection techniques or official manipulation), or they may be absent altogether. The absence of critical socio-economic and demographic data, which is essential in so many qualitative analyses, also impedes the effective execution of established survey methodologies (i.e., Primary Sampling Units stratified by population size).

Sensitive Subjects

Surveyors report gaps in responses for certain sensitive subjects such as participation in informal economies, as well as “taboo” subjects such as sexual preference, drug consumption, or lifestyle behavior. They also report scarce responses for questions where there may be a fear of government retribution, e.g., perceptions of corruption and legitimacy of government. Even where respondents are guaranteed anonymity, surveyors report reluctance to provide answers to questions which respondents feel may be used against them by their governments.

- Baseline data
- Recent data
- Census data in many countries
- Sensitive subjects

Distinct from gaps in qualitative data are the challenges associated with collecting and analyzing qualitative data. In terms of collection, identifying the most appropriate unit of analysis is a problem that has befallen many quantitative and qualitative data collection efforts in ethnically diverse regions. For populations organized around governmental units like provinces, counties, districts, cities, and villages, generating samples from these units is straightforward and applicable, since administrative units or grid squares may be used. In ethnically diverse societies, however, the composition, size, and location of groups change over time, which complicates sample frames. Because nomadic populations move fluidly throughout a region, sampling from government administrative units would not capture the same population over time and would most likely produce biased (and potentially misleading) results.

MS&T designers have also expressed some frustration over the need to combine datasets or to integrate new data with existing datasets, since the lack of a uniform collection methodology makes this difficult. There are many cases where two or more datasets track the same variable but use different methodologies, or where recent data collection efforts could

augment existing datasets. Combining disparate data streams is a worthwhile practice in order to produce a comprehensive dataset. Yet doing so is tedious and time-consuming – requiring analysts to “clean” the data and account for differences in collection methodologies to maintain the integrity of the data.

Collecting data in Africa can also be impeded by the high cost of operating on the continent and limited access of Americans into traditional communities. For this reason, existing collection efforts typically hire local firms to perform these duties, since they are more cost-effective than American researchers and are culturally sensitized to local conditions. Interviewees note, however, that quality control can become a time- and resource-consuming burden in these cases. One respondent from a large survey organization noted he spends half of his time in Africa overseeing quality control measures and directs an entire team in Nairobi whose sole function is to travel throughout the continent and ensure adherence to “best practices” across data collection teams. Nonetheless, this expenditure of time and resources is well worth the investment in order to collect the most data possible.

IDA | Challenges of Collecting/Analyzing QD from Africa

- Unit of analysis
- Lack of a uniform data collection methodology
- High cost and limited access of U.S. data collectors

The bureaucracy and corruption that typically accompany research in Africa are commonly-cited impediments to freedom of movement within a country and therefore effective data collection. Obtaining permission from government and traditional authorities to conduct surveys can be complicated and time-consuming. This is another reason that contracting with local partners who typically already have the necessary permission to conduct research is a common practice.

Several survey organizations reported some cultural bias when using the Likert scale.¹² For example, on a 5-point scale with 1 being “lowest” and 5 being “highest,” some respondents, based on cultural predisposition, treat only responses 2 through 4 as available options, while others view 1, 3, and 5 as the only ones available. It is possible to overcome these biases by assessing changes over time (i.e., the trends) and making as many of the questions in the “yes” or “no” (i.e., dichotomous) format.

Another challenge for researchers in Africa is ensuring that they capture significant and relevant data on emerging salient issues before receiving cues that the issue is an important one. “Getting ahead of the curve” is critical when examining emerging trends, instances of major social change, and potential security threats. Phase II of this study will address this issue by soliciting inputs from local Africans with far better insight into the conditions on the ground than U.S.-based researchers.

An important difficulty that was reported to IDA is the occasional misuse of Measures of Performance (MOPs) for Measures of Effectiveness (MOEs). Performance indicators usually take the form of quantitative values that allow researchers to determine whether performance requirements have been met. MOEs typically take the form of qualitative indicators that provide an assessment of the impact of the performance measures. In the absence of qualitative data to measure MOEs, analysts have relied on MOPs – for example, using the number of free media outlets (an MOP) to describe the effectiveness of national governance, where local attitudes and perceptions of legitimacy (qualitative data) would have far more explanatory power. This misuse of MOPs for MOEs presents a challenge in the final stages of program evaluation where policy-makers wish to measure the return on their investments in terms of effectiveness.

¹² The Likert scale is the most widely used approach to survey research where responses are chosen among a ranking of multiple categories.

- Bureaucracy and corruption
- Culturally determined response bias
- Capturing emerging issues
- Conflation of Measures of Performance and Measures of Effectiveness

IDA is currently underway with Phase II of this research – engaging African partners from various research institutions. Thus far, African researchers have been eager to provide their inputs to this study and appreciate the opportunity to do so. With their insights, IDA is beginning to compile a list of data points that are significantly different from those identified by USG M&S designers. As a result, IDA is encouraged by their participation and confident their insights will contribute to a QDCS that is fresh and relevant.

IDA is also beginning to analyze the components necessary for a broad strategy as described in Phase III. Based on findings

so far, IDA sees three distinct areas where the USG can focus its efforts to fill some major qualitative gaps:

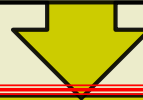
- Immediate improvements, such as the use and sharing of new qualitative data sources.
- A near-term “surge” in data collection for the most frequently cited data gaps.
- A long-term plan to ensure a sustainable stream of required data, including technical capacity building of African partners.

Phase I: “Gap” analysis of MS&T relating to Africa

Reliability: Is data available? On a consistent basis? Only some of the time? Does it consistently capture the same information?

Validity: Is data accurate? Does it reflect the concept it purports to?

Structure: Is the data structured in a format conducive for use in M&S?

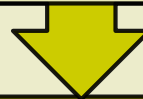


Phase II: Solicit input from African partners

Are gaps identified by Americans the only ones? Are they relevant?

Identify additional data points from African perspective

Including Africans as active participants in data collection efforts ensures that data is as accurate and as valid as possible



Phase III: Strategy Development

Design a Qualitative Data Collection Strategy (QDCS) for acquiring missing data relevant to the study and analysis of Africa

Appendix: MS&T Survey

Name: Competitive Influence Game (CIG)

Producer: Johns Hopkins University, Applied Physics Lab (APL)

Type: Simulation (Independent & Federated) – CIG is an “independent” simulation because it can run entirely on its own when provided sufficient amounts of data inputs. It also has the ability to federate (i.e., the ability to combine with multiple model or simulation inputs), as it is equipped with a Federation Object Model (FOM), which describes the shared object, attributes and interactions for the whole federation. It is unclear at this time whether the CIG FOM satisfies governmental, High Level Architecture (HLA) standards.

Purpose: Currently used to support exercises and high-level wargaming (e.g., the AOWG/AWG Cycles), its developers at APL originally conceived of it as an attempt to provide a generalized behavioral model characterized after the fictional Seldon equations (the one elaborated upon by Isaac Asimov in the 1951 novel, *The Foundation*). Asimov described the Seldon equations as essentially statistical models with historical data of a sufficient size and variability that they are collectively

representative of the population under consideration. The intent is not to provide point predictions that accurately capture the behavior of an individual but instead to generate accurate forecasts of how populations will behave in the aggregate. CIG adheres to the spirit of Seldon equations in structure but variation in the number, quality, and empirical anchoring of inputs causes it to differ in form.

Inputs: Generation of behavioral outcomes in CIG is similar to that of tabletop board games, such as Risk and others that model probabilistic outcomes using die rolls. Although probability distributions are always normal or “bell curves,” their shape (i.e., location of mean values and population variance) results from the conditional mapping of behavioral outcomes within the game. Currently, the setting of “initial conditions” or starting values for data in the simulation along with the properties governing values for the conditional mappings occurs primarily according to subjective inputs from SMEs. While all of the SME-elicited relational estimates are qualitative, the nature of “initial conditions” inputs describing existing conditions varies between quantitative and qualitative.

Name: Composite Vulnerability Map

Producer: University of Texas, Climate Change and African Political Stability Program (CCAPS)

Type: Web-Based Tool

Purpose: The Composite Vulnerability Map models which parts of Africa are most vulnerable to climate change in the mid 21st century range. It provides scholars, policymakers, analysts, and those supporting them with the ability to visualize imagery, events (from human behavior), and other types of related data in the effort to characterize the relationship between various physical and social environmental variables and human conflict. Its most mature capability is the ability to generate layered visualizations containing imagery data, such as precipitation, and a large variety of violent events (i.e., sub-nationals against sub-nationals, states against sub-nationals, and sub-nationals against the state). Data on governance characteristics will eventually extend beyond that available in other datasets (e.g., PolityIV) to

describe state features of constitutional processes and other manner of non-quantitative and subjective information. Besides making visualization tools accessible by the public, the project also provides links for downloading the represented data.

Inputs: Imagery data (e.g., drawn from NASA, NGA, and other similar sources) and originally collected, spatio-temporal (i.e., geo-located and time-coded) event data from systematically coded news events. Maintaining updated imagery information is an external matter for the project and characterizing historical processes leading to socio-political events, such as referenda and drafting, are fixed, historical features of countries requiring only one pass to provide information (unless the feature in question changes). On the other hand, event data in the tool suffers from a lag between social processes generating events on a daily/weekly/monthly/quarterly rate (depending on the nature of conflict in the specific locale) and the ability to code them into datasets.

Name: Cultural Geography¹³

Producer: United States Training and Doctrine Command (TRADOC), Analysis Center (TRAC)-Monterey

Type: Pseudo-Agent Based Model (ABM)

Purpose: The purpose of CG is to provide a platform for considering the consequences of kinetic and non-kinetic actions taken by military actors within simulated socio-cultural environments. It is part of the Social Impact Model (SIM) system, which is a type of model federation described as “a tool for irregular warfare adjudication, analysis, and validation.” Given that the capability hails from TRADOC, its primary purpose is to support training in areas such as the selection and prioritization of courses of action (COAs) within the context of a COIN socio-cultural environment.

Inputs: CG possesses the ability to model micro-level agents, but the complexity of its architecture and vastness of its parameters has in practice led to the modeling of “representative agents.” Examples of such actor agents include a community, a government, an ethnic group, an insurgent, and so forth. Individually, requisite inputs include data on the preferences these actors hold over a variety of outcomes, prior beliefs about the preferences of other actors, relational mappings for actions and changes to the environments as indirect influences on outcome evaluations, and so forth. Social network components in the model require data on the relationships between groups, i.e.,

who shares a connection with whom and the relative value of this relationship. These are just some of the numerous data inputs for calibrating parameters in the model.

¹³ The IDA team received access to the code and other documentation for CG. Additionally, IDA coordinated with National Defense University which is overseeing a validation project for Cultural Geography and ATHENA. Working through the complex architecture and processes of CG is an extensive effort extending beyond the scope of IDA’s tasking, so the team has relied upon available documentation as well as interviews with TRAC-Monterey and NDU to complete this entry in the report.

Name: Geospatial Information Awareness/Infection Disease (GIA/ID)

Producer: Naval Research Lab (NRL)

Type: Computational Analytic Model

Purpose: Africa is a continent where the emergence and spread of disease are persistent threats. Enhancing geospatial information for the purpose of situational awareness has gained traction and considerable development throughout the West. GIA/ID is an initiative led by NRL to expand the community of interest and practice throughout Africa. As an initial step, GIA/ID is a “proof-of-concept” attempt to demonstrate the ability to identify the emergent flash point of a disease (geo-referenced), to track its spread (geographically and temporally), and to identify factors—including social and environmental – associated with these empirical trends. The hope is that if conducted successfully, analysis of these three components will provide indicators and warnings for American and partnering forces. Additionally, outputs from GIA/ID should identify interventions tailored to the specific socio-environmental conditions responsible for identified pandemics, limiting the need to rely upon “cookie cutter” solutions commonly applied under conditions characterized by low information.

Inputs: Current inputs to GIA/ID include an extensive surveying of the population in the Sierra Leone town of Bo, used to establish what NRL analysts described as the denominator. Specifically, the denominator is a geo-referenced count of the population on a grid-by-grid basis across the territory. This required extensive resources to collect. Another input is the counting of diseased individuals, which constitutes the numerator. At the time IDA discussed the project with NRL, the identification of cases was relatively accurate (i.e., use of a university-donated, genomic analyzer facilitated the efficient identification of pathogens in blood serum), as too was its temporal tagging (i.e., association of the identified case with a date of collection – though there is a difference between identifying when transmission of a pathogen took place versus when a patient makes it to a clinic or hospital). What the data lacked was an implemented means to geo-reference the reported incidence of disease within the grids established during the initial surveying of the population. Territory in Bo is not systematically organized in a manner that residents can readily provide meaningful addresses, which was the primary culprit for this initial lack of geo-referenced cases. A proposed solution at the time of the interview included having doctors present maps of the area to patients for them to use when identifying their place of residence.

Name: HOA-Viewer

Producer: Department of State (DoS), Humanitarian Information Unit (HIU)

Type: Web-Based Tool

Purpose: Intentions for the HOA-Viewer are twofold. First, HIU wants the tool to equip users (e.g., analysts, service providers, policymakers, and so forth) with the ability to visualize and interact with data in a manner that exploits geospatial and temporal characteristics of humanitarian crises (both the crisis events themselves as well as the circumstances preceding and following them). HIU also aspires for HOA-Viewer to be an

analytic support tool by eventually infusing it with qualitative and quantitative methodological functions.

Inputs: HOA-Viewer inputs include a broad array of imagery data (theoretically, the system can capture any level of imagery data available), United Nations Humanitarian Crisis (UNHCR) Reports (unstructured text), and other geospatial data (e.g., ethnicity and population size polygons as well as event point data). Metadata each input includes geospatial and temporal components, which enable the viewer to visualize on maps various patterns of events (currently, the focus is on the representation of climate imagery data).

Name: Information Velocity 2.0 (IV2)

Producer: Office of the Secretary of Defense, Science and Technology

Type: Web Information Harvesting Tool

Purpose: Surveying populations is an effective means for tracking attitudes and sentiment, but it is a timely process with uncertainty surrounding the conditions producing responses. Rather than survey populations directly, Web 2.0 products, such as Twitter and Facebook, provide the opportunity to track attitudes and sentiments in a populous, as expressed directly by individuals (i.e., without the response and construction biases of surveys but also without their controllability). IV2, which is currently under development as a governmental specification for currently, “commercial off the shelf” (COTS) products, plans to tap into this resource in the effort to provide AFRICOM (and by extension other global combatant commands) with the ability to track and potentially predict the occurrence of flash points associated with mass unrest throughout the African area of operations. [IV2 and similar capabilities under development, such as Mitre’s Social Radar, use the examples of the London riots and the Arab Spring as cases in point for harnessing Web 2.0 technologies]. IV2 developers envision that automated reference extractions from Web 2.0 associated with Web 1.0

(e.g., newsfeeds along with company and individual profile webpages among others) will result in a broader contextual understanding, higher situational awareness, and potential ability to act than either capability alone provides.

Inputs: IV2 inputs will include Web 2.0 (e.g., Twitter, and Facebook) feeds in addition to Web 1.0 targeted page scraping, conditioned on Web 2.0 extractions. Importantly, when thinking about the application of IV2 and similar technologies, it is important to consider the informational austerity of the population in question and the targeted objective of the capability. Public opinion polling, which – when done well (e.g., according to standards followed by AfroBarometer, Gallup, and the State Department Office of Opinion Research among others) – is representative of the population in question with an identifiable degree of uncertainty (i.e., with confidence intervals on reported percentages). If the goal is to use the IV2 capability as an alternative to public opinion polling, then it will be necessary to use it on online populations that are accurate subsets of the entire population (i.e., randomly available online in a manner similar to samples generated from randomized, stratified sampling used to construct survey populations) or to at least have determined the systematic bias distinguishing expressed online sentiments from those counterfactually gathered in person.

Name: RiftLand

Producer: Center for Social Complexity, George Mason University

Type: Agent Based Model (ABM)

Purpose: Generally speaking, RiftLand models humanitarian crises in East Africa. Based on the description of its predecessor, RebeLand, the analytic goal of the model is to study conditions of political stability, specifically the ability of a system to withstand, various forms of stress, such as social, economic, political, or environmental. The name of the model implies that it focuses on the area in Kenya known as the Rift Valley. Following the 2007 Presidential elections, the Rift Valley was one of the areas that erupted into violence as disputed election results resonated with a long history of inter-ethnic rivalry and conflict among residents. Numerous violent events and large-scale internal displacement resulted in widespread instability throughout the Rift region. IDA infers that one goal of RiftLand is to identify regional or functional areas where government action may help to prevent future instability.

Inputs: RiftLand, as a “real world” version of RebeLand, is an attempt at modeling an entire polity. According to documentation for RebeLand, some of the basic inputs required for doing this include a range of geospatial information (e.g., provincial boundaries, topography and land cover, location and size of cities, location as well as type and amount of natural resources), location along with type and composition of military (state and non-state) forces, climate data (e.g., rainfall/drought, wind, and temperature), hydrology, and so forth.¹⁴ Corresponding data requirements for RiftLand, beyond the basic descriptive

characterizations of the local population, are not yet documented for public consumption.

Modeling societal effects of naturally-occurring or manmade phenomena require values for those actions as well as data on the relational mapping between changes in these values and outcomes of interest. Other implicit inputs to RebeLand include how changes in community context and individual wellbeing affect recruitment of rebel and other anti-state groups. Authors emphasize the characterization of community issues relative to government activity. Abstractly, it is possible to work through the analysis of this problem without “real world” data, but linking the two (i.e., determining what the definition of an issue and its relevant dimensions are for coding in a dataset for ingestion to the model) is necessary for accurate modeling. Documentation for RebeLand does not explicitly identify contextual and relational data as necessary inputs, but it is clear that the utility of RiftLand depends upon capturing this information along with the descriptive data already identified as inputs.

¹⁴ Claudio Cioffi-Revilla and Mark Rouleau, “MASON RebeLand: An Agent-Based Model of Politics, Environment, and Insurgency” *International Studies Review* 12, 31-52, 2010.

Name: Unnamed

Producer: Naval Postgraduate School (NPS), Operations Research Department

Type: Web-Based Data Visualization Tool (with future possibilities for analysis development)

Purpose: This tool under development at NPS intends to make survey data more accessible to end-users who are not well versed in the handling and exploitation of survey data. Currently, exploitation of raw, survey data requires some facility with software tools, such as those in the Microsoft Office suite (Excel and Access) or more traditional statistical analysis platforms (e.g., R, Stata, SPSS, and Gauss to name a few). Even those capable of using such programs find it difficult to visualize and understand calculated results geographically, because doing so

necessitates the additional skills required to work either mapping functionalities within the aforementioned platforms (mainly the alternate packages available in R) or to import and manipulate them within a geospatial analysis platform, such as Esri's ArcGIS suite. The product under development at NPS seeks to overcome both hurdles for end-users who do not have time to develop the necessary skill sets but nonetheless need the data and the insights it brings.

Inputs: The tool ingests survey data, which makes the quality of its outputs entirely dependent upon that of its inputs. This means it is sensitive to common survey data issues, such as sample construction, question validity, timeliness, along with a host of others. Efforts made to resolve these problems will translate directly into the quality of insights the NPS visualization tool provides.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) June 2012		2. REPORT TYPE IAD Final		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Designing a Qualitative Data Collection Strategy (QDCS) Africa -- Phase I: A Gap Analysis of Existing Models, Simulations, and Tools Relating to Africa				5a. CONTRACT NUMBER DASW01-04-C-0003	
				5b. GRANT NUMBER — — —	
				5c. PROGRAM ELEMENT NUMBER — — —	
6. AUTHOR(S) Ashley N. Bybee, Dominick E. Wright				5d. PROJECT NUMBER — — —	
				5e. TASK NUMBER AI-55-3061.0.0	
				5f. WORK UNIT NUMBER — — —	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, Virginia 22311-1882				8. PERFORMING ORGANIZATION REPORT NUMBER D-4629 H12-000748	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Glenn A. Fogg Director, Rapid Reaction Technology Office (703) 746-1343 2231 Crystal Drive, Suite 900 Arlington, VA 22202				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) — — —	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited. Director, Rapid Reaction Technology Office. 28-08-2012.					
13. SUPPLEMENTARY NOTES — — —					
14. ABSTRACT This document summarizes the findings of the IDA's survey of models, simulations, and relevant tools currently being used to analyze the African continent. It identifies the most pressing gaps in data and captures the most significant challenges or obstacles to effective data collection. This phase of research supports the final objective, which is to draft a Qualitative Data Collection Strategy (QDCS) that addresses "gaps" where insufficient qualitative data exist for the African continent					
15. SUBJECT TERMS Africa, qualitative, data, social science, models, simulation					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Unlimited	18. NUMBER OF PAGES 44	19a. NAME OF RESPONSIBLE PERSON Ashley Bybee
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code) 703-845-2288

